RotorNet: A Scalable, Low-complexity, Optical Datacenter Network

William "Max" Mellette

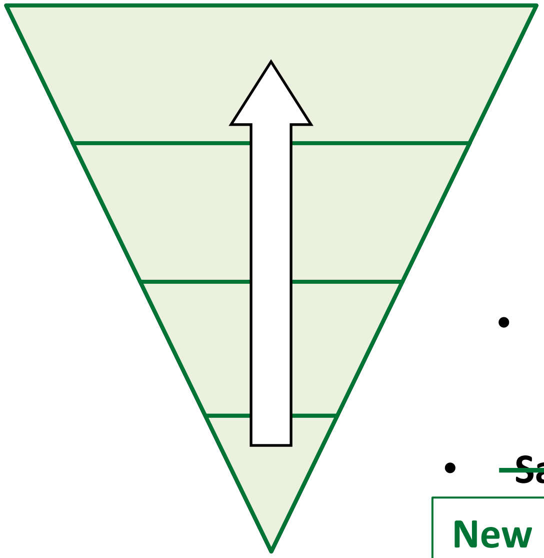
Rob McGuinness, Arjun Roy, Alex Forencich,
George Papen, Alex C. Snoeren, and George Porter

UC San Diego

Toward 100+ Petabit/second datacenters

Challenge: deliver (very) low-cost bandwidth at scale

Co-design:
Protocol
Topology
Hardware



- **New protocols**
  Load balancing, congestion control, …
- **New topologies**
  Jellyfish, Longhop, Slimfly, …
- **New hardware**
  Optical circuit switching, RF/optical wireless, …
- ~~**Same switching model**~~

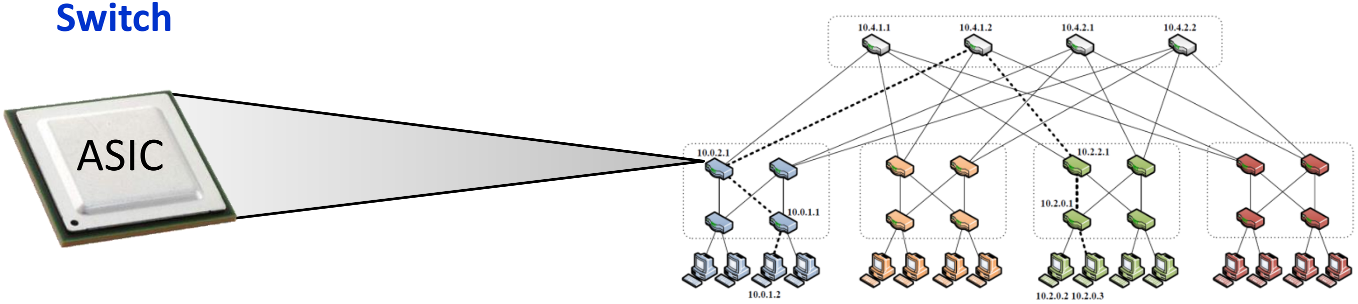
New "Rotor" switching model

RotorNet → "Future-proof" bandwidth (2× today) + simple control + …

Don’t packet switches work fine?

Electronic Packet Switch

Fat Tree (Sigcomm ‘08)



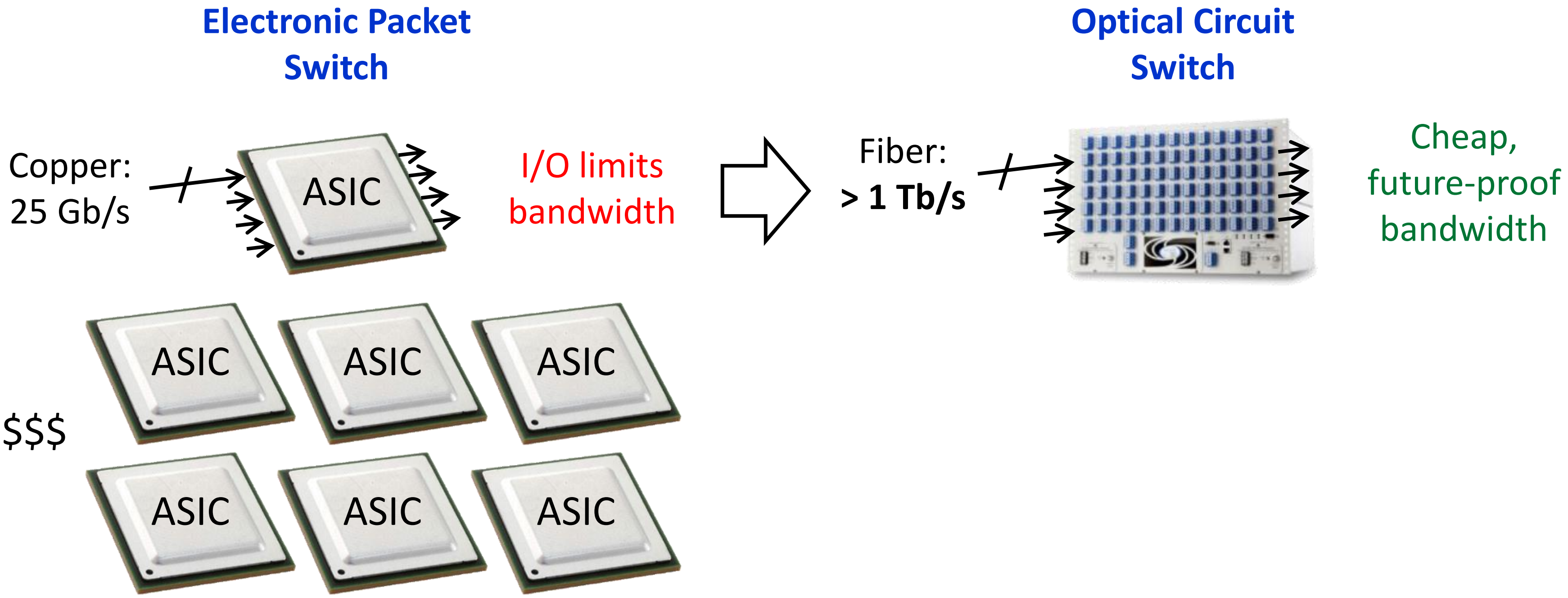
Packet switch capacity growth:
~ 2× / 2 years

<

Network capacity growth:
~ 2× / year

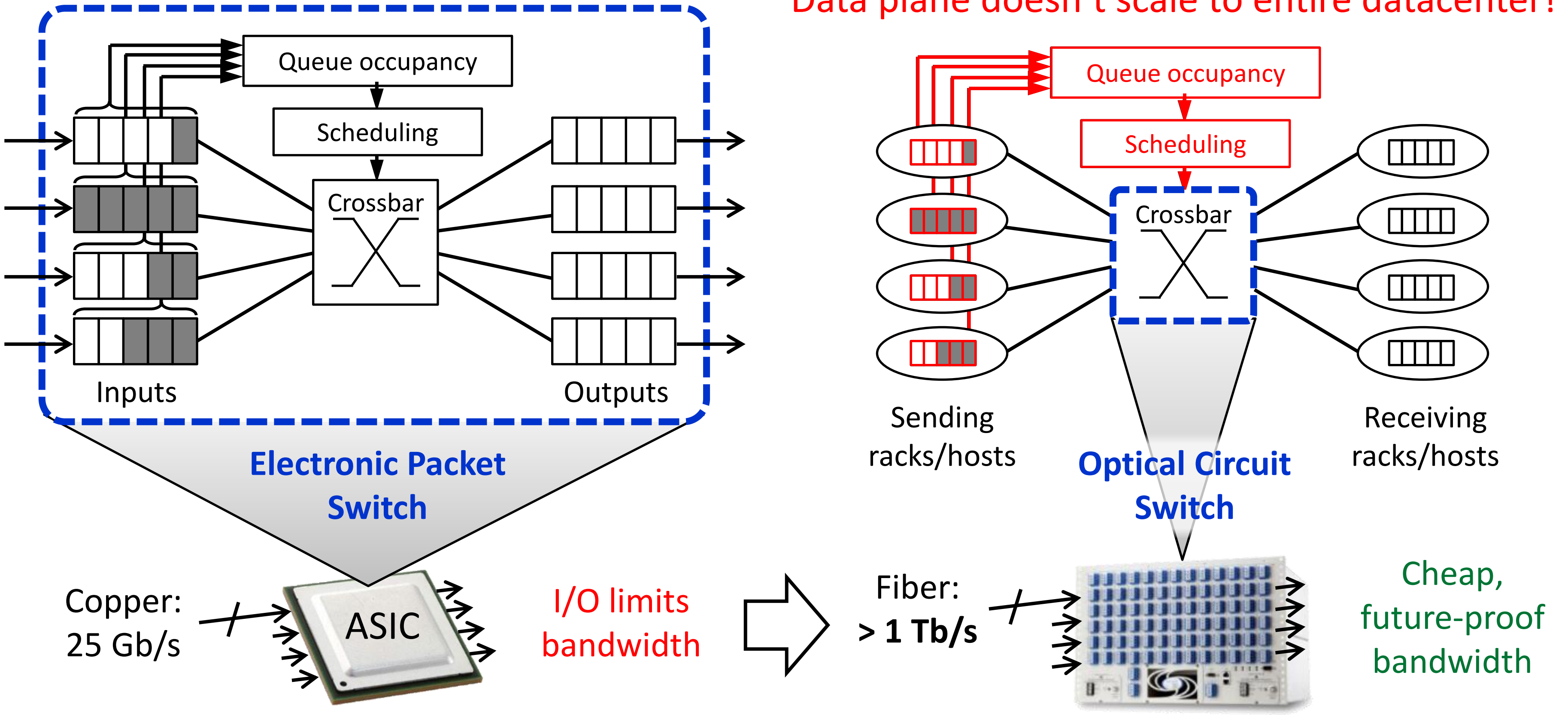
(A. Singh et al., SIGCOMM 2015)

Optical switching – benefits & barriers



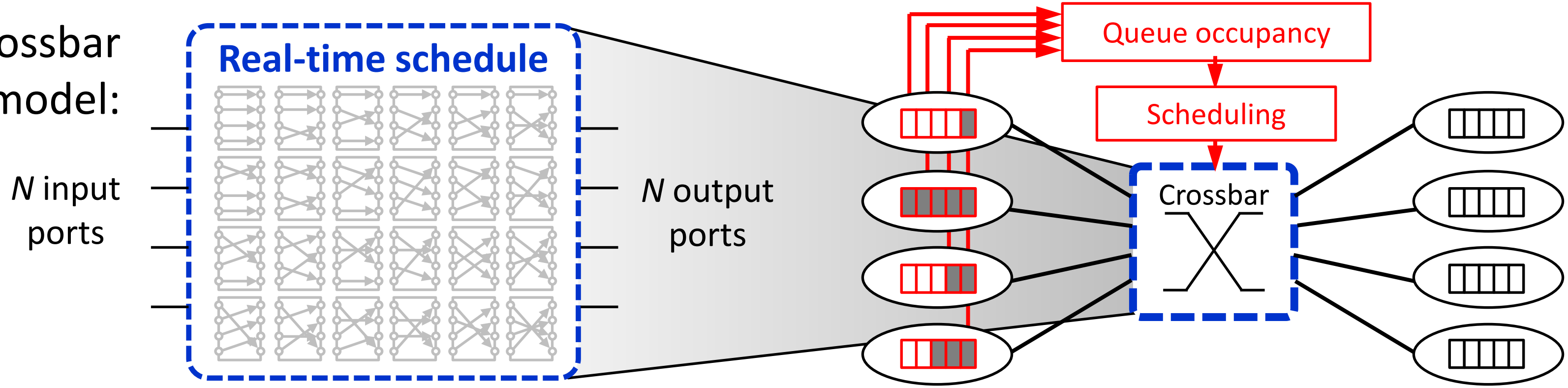
Optical switching – benefits & barriers

Data plane doesn't scale to entire datacenter!



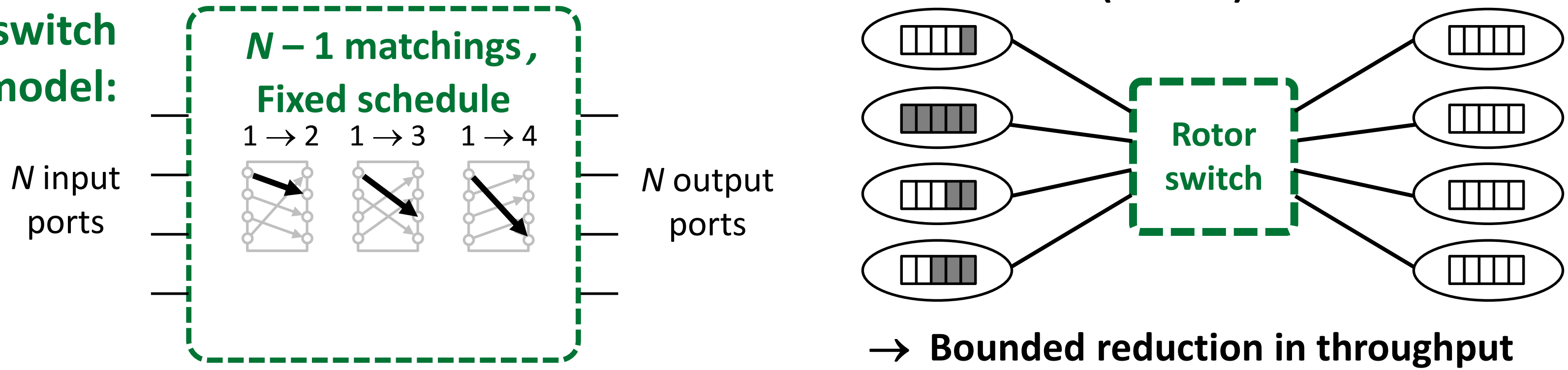
Rotor switching model simplifies control

Crossbar model:



→ No (central) control

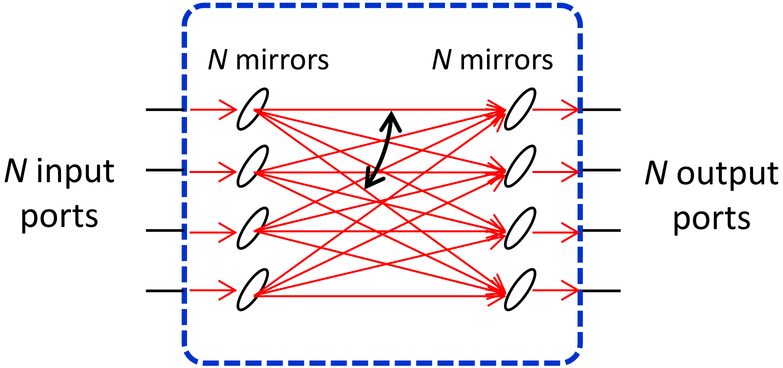
Rotor switch model:



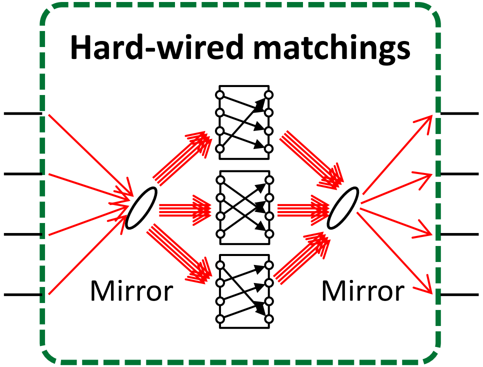
→ Bounded reduction in throughput

Rotor switches have a simpler implementation

Optical Crossbar:



Optical Rotor switch:



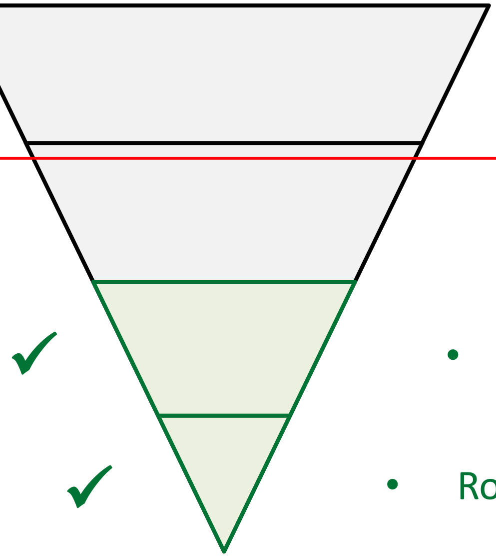
- Cost and complexity scale with:

Ports

Matchings (<< Ports)

Ex. 2,048 ports: 4,096 mirrors, 2,048 directions | 2 mirrors, 16 directions

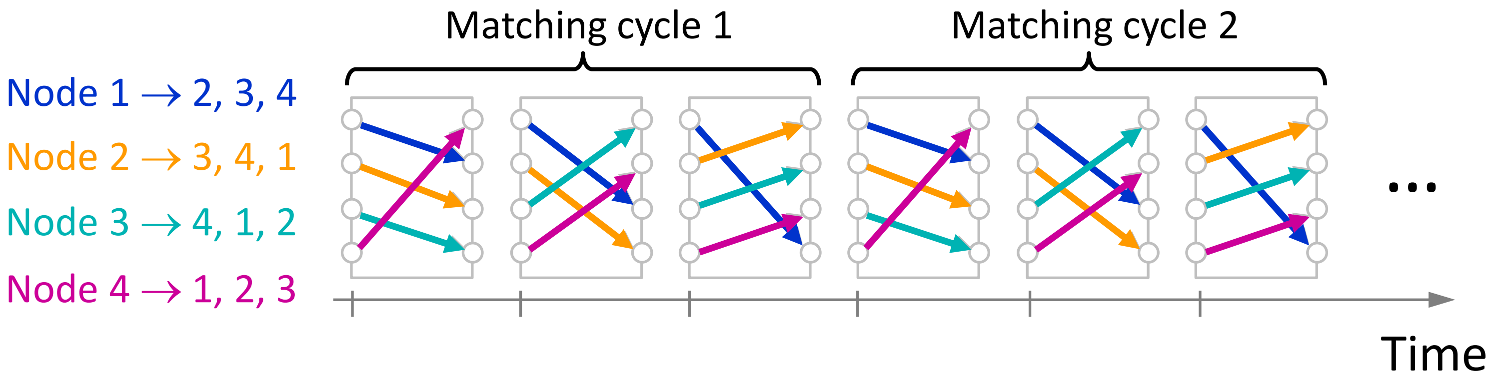
RotorNet architecture overview



- **Forwarding?**
- **Topology?**
- Optical Rotor switch → **More scalable**
- Rotor switching model → **Simpler control**

1-hop forwarding over Rotor switch

- Wait for direct path:

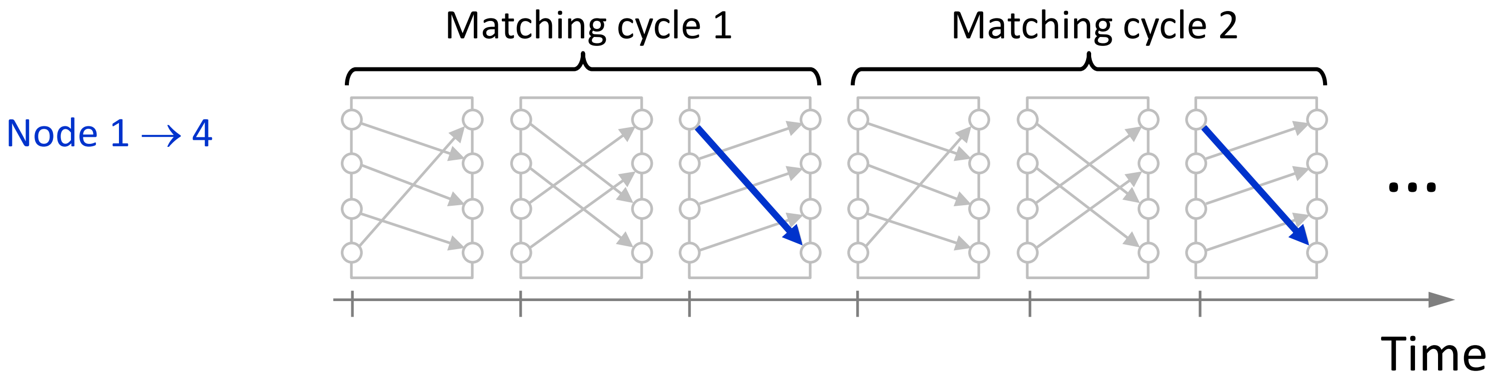


Uniform traffic → 100% throughput

- But datacenter traffic can be sparse …

1-hop forwarding & sparse traffic = low throughput

- Wait for direct path:



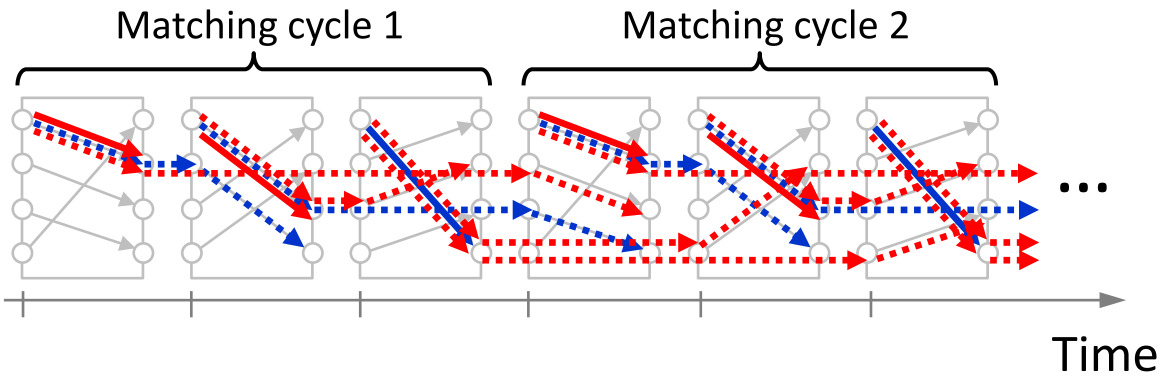
Problem: single flow → 33% throughput

- Hint at improvement: network is underutilized

2-hop forwarding better for sparse traffic

- Not new: Valiant ('82) & Chang et al. ('02)

Node 1 → 4, 3, 2
Node 2 → 3, 4, 1
Node 3 → 4, 1, 2
Node 4 → 1, 2, 3



Throughput: **Single flow** **33% (1-hop) → 100% (2-hop)**

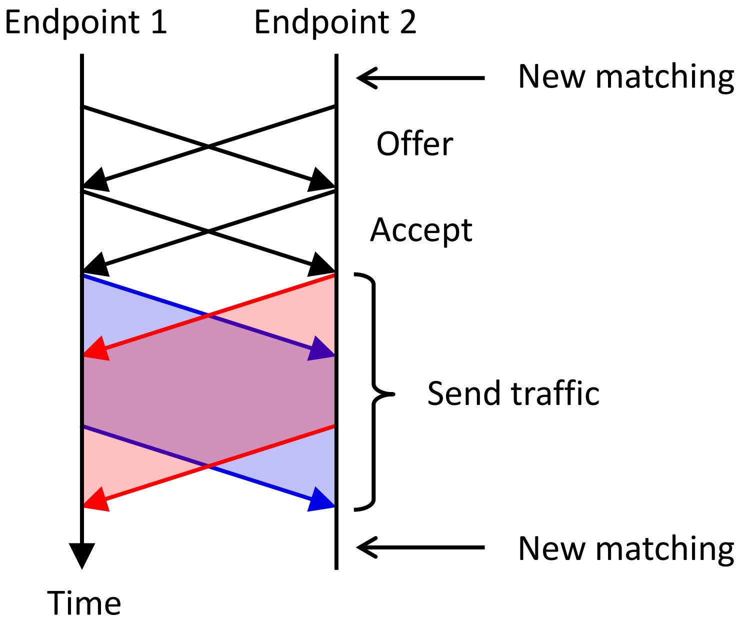
Uniform traffic **100% (1-hop) → 50% (2-hop)**

- Optimization: can we adapt between **1-hop** and **2-hop** forwarding?

RotorLB: adapting between 1 & 2-hop forwarding

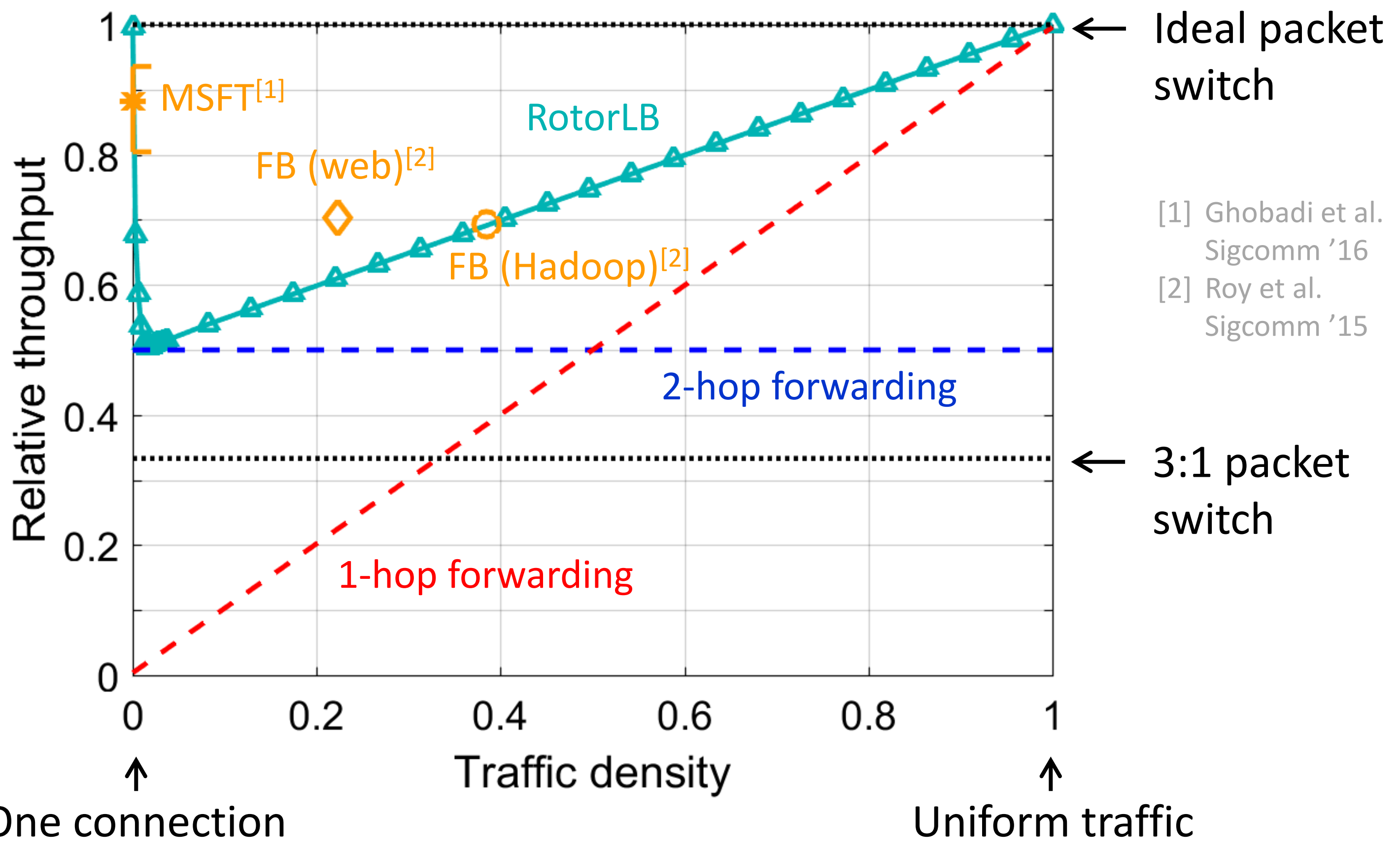
RotorLB (Load Balancing) overview:

- Default to 1-hop forwarding
- Send traffic over 2 hops only when there is extra capacity
- Discover capacity using in-band pairwise protocol:



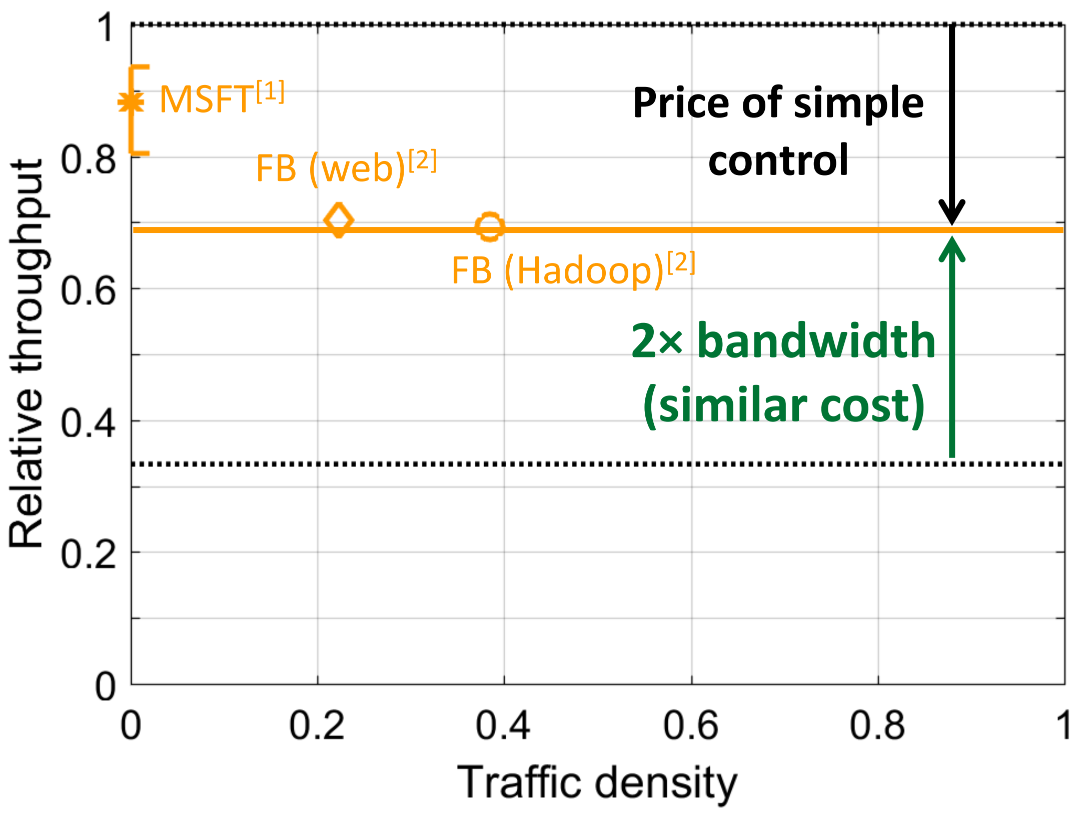
→ **RotorLB is fully distributed**

Throughput of forwarding approaches (256 ports)



[1] Ghobadi et al. Sigcomm '16
[2] Roy et al. Sigcomm '15

Throughput of forwarding approaches (256 ports)

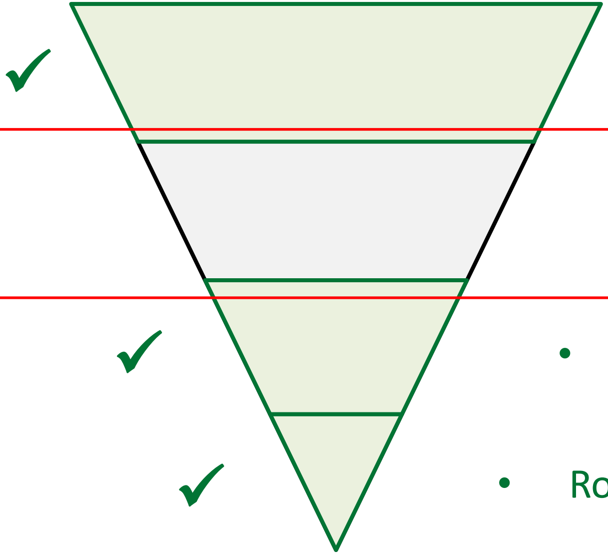


← Ideal packet switch

← 3:1 packet switch

[1] Ghobadi et al. Sigcomm '16
[2] Roy et al. Sigcomm '15

RotorNet architecture overview

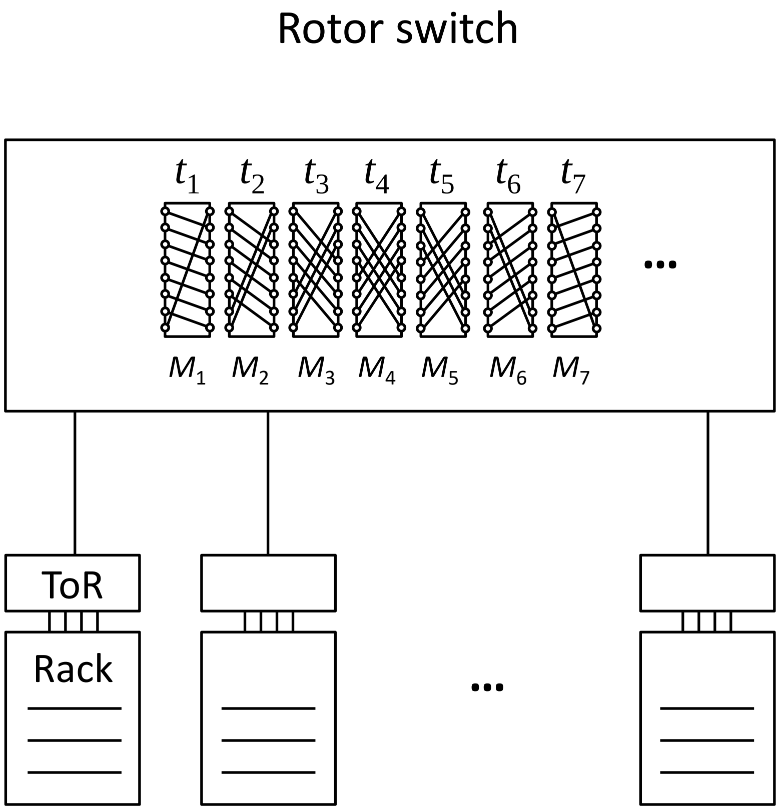


- RotorLB → **Distributed, high throughput**
- **Topology?**
- Optical Rotor switch → **More scalable**
- Rotor switching model → **Simpler control**

How should we build a network from Rotor switches?

At large scale:

- **High latency:** Sequentially step through many matchings
- **Fabrication challenge:** Monolithic Rotor switch with many matchings
- **Single point of failure**



Distributing Rotor matchings = lower latency

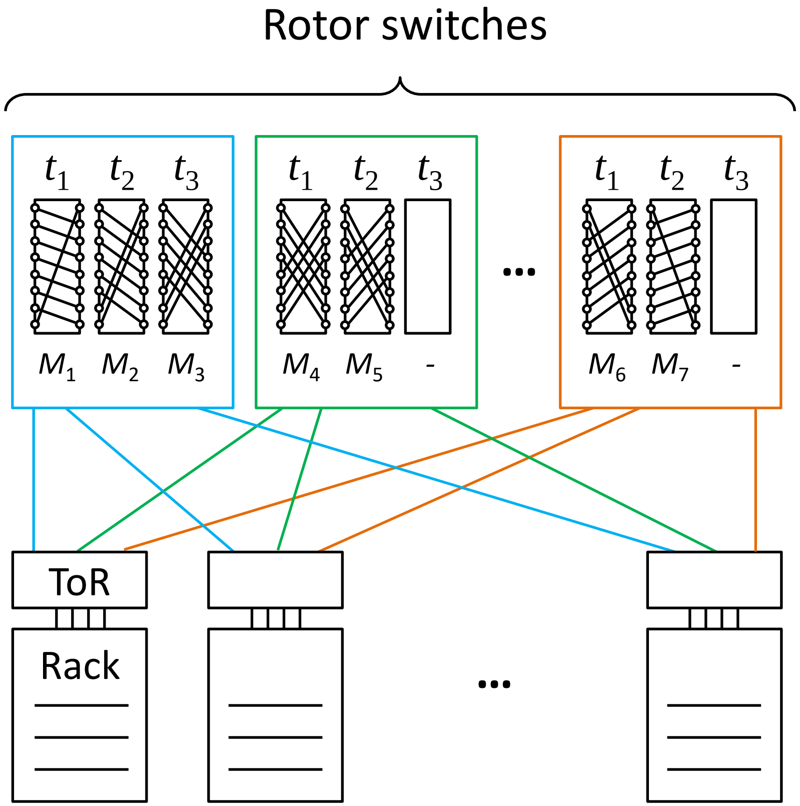
Fault tolerant

Reduced latency:

- Access matchings in parallel

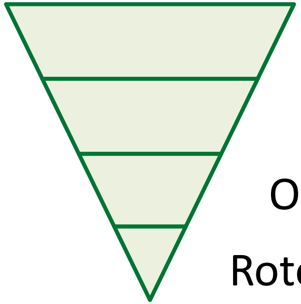
Simplifies Rotor switches:

- Matchings << ports
- More scalable, less expensive



Rotor switching is feasible today

Validated feasibility of entire architecture:
(8 endpoints)



RotorLB

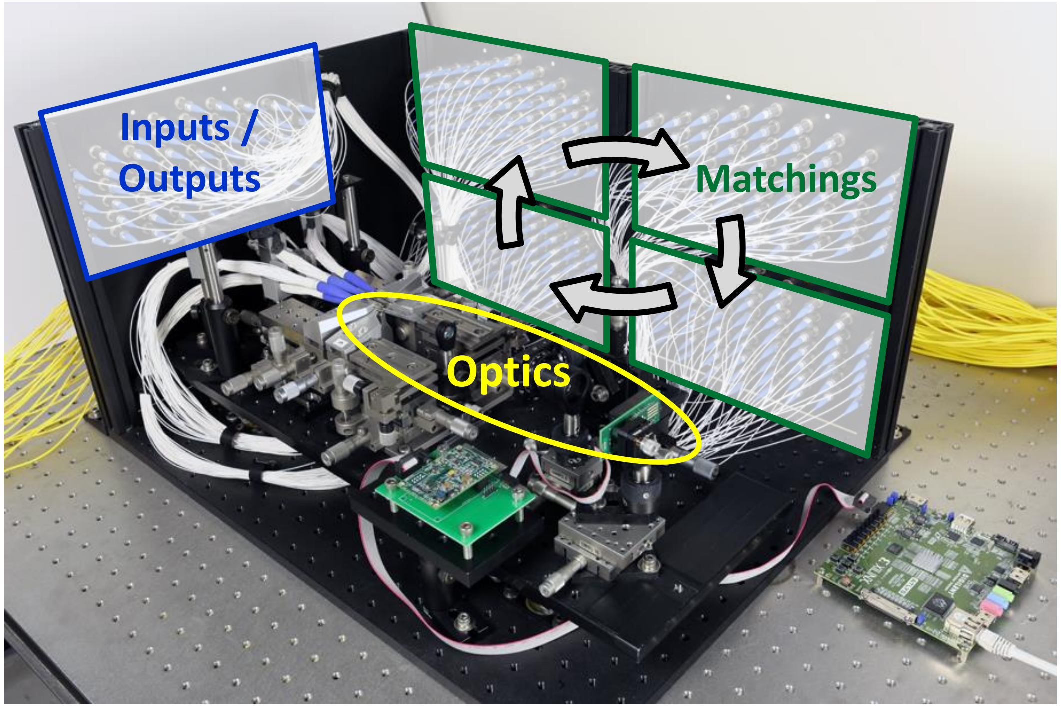
RotorNet topology

Optical Rotor switch

Rotor switch model

100× faster switching than crossbar

Prototype Rotor switch



RotorNet scales to 1,000s of racks

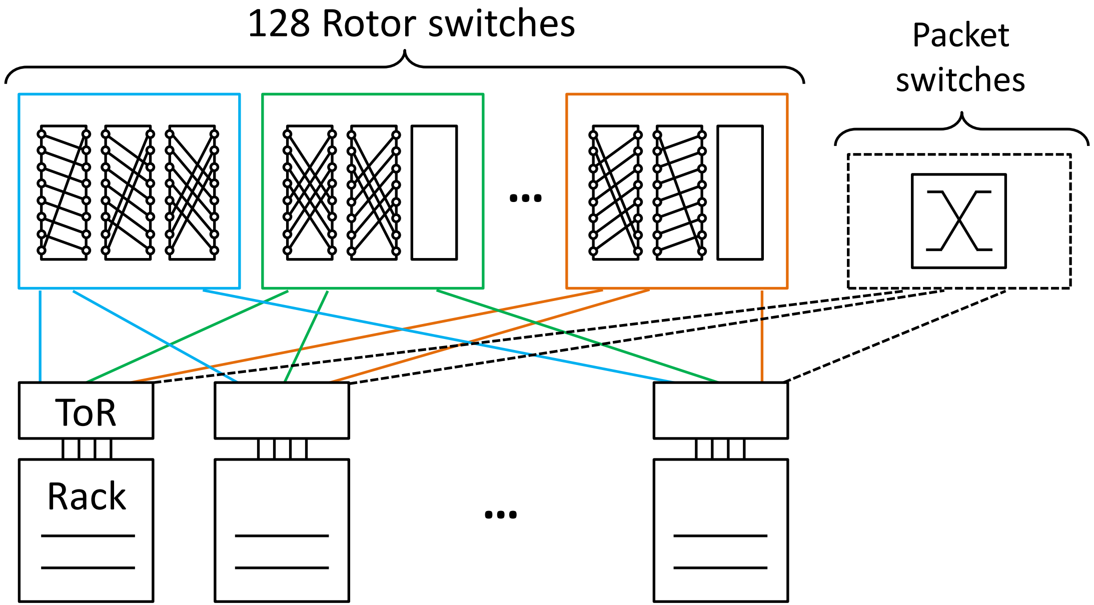
- Rotor switch design point: 2,048 ports, 1,000× faster switching than crossbar

  Details in: W. Mellette et al., *Journal of Lightwave Technology* ’16
  W. Mellette et al., *OFC* ’16

- 2,048-rack data center:
  → **Latency (cycle time) = 3.2 ms**

- Faster than 10 ms crossbar <u>reconfiguration time</u>

- Hybrid network for low-latency applications



RotorNet component comparison

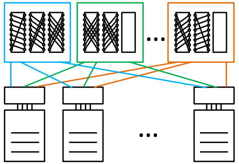
Network	# Packet switches	# Transceivers	# Rotor switches	Bandwidth
3:1 Fat Tree	2.6 k	103 k	0	**33 %**
RotorNet, 10% packet	2.3 k	84 k	128	**70 %**
RotorNet, 20% packet	2.5 k	96 k	128	**70 %**

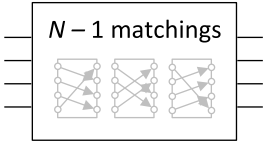
RotorNet delivers:

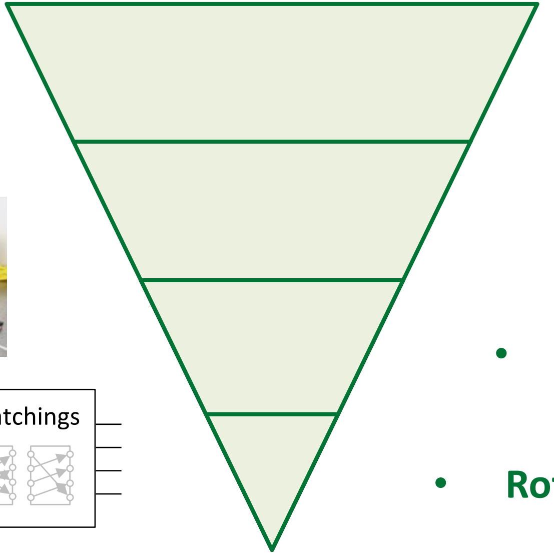
- Today: Bandwidth 2× less expensive
- Future: Cost advantage grows with bandwidth
- **Benefits of optical switching without control complexity**

A scalable, low-complexity optical datacenter network

RotorNet architecture:







- **RotorLB** → Distributed, high throughput
- **RotorNet topology** → Fast cycle time
- **Optical Rotor switch** → More scalable
- **Rotor switching model** → Simpler control

This work was supported by the NSF and a gift from Facebook